Copernicus

Datasøgning

Kristian Søgaard og Morten Kaaber

Gatehouse

March 14, 2021

Bullet Points

- Hvad er udfordringen?
- Hvilke programmer/metoder har vi brugt?
- Hvad er der af begrænsninger i det hentede data?
- Clustering fremgangsmåde
- Clustering ved brug af Jaccard distance

Hvad er udfordringen?

- Store mængder data er frit tilgængeligt på Copernicus.eu.
- Dog er det på nuværende tidspunkt besværligt at finde den data, man skal bruge.
- Strukturen på Copernicus.eu:
  - 6 overkategorier(Atmosphere, Climate, Emergency, Land, Maritime, Security).
  - Ydereligere underforgreninger i hver overkategorier.
  - Alle overkategorier har sin egen struktur.
  - Dataene i hver kategori er ikke centraliseret 1 sted.
- For at gøre det nemmere at finde den data en eventuelt bruger er interesseret i vil vi implementere en search engine.

Eksempel: Udsnit af placering af forskellige dataset i Atmosphere kategorien

Region	Dataset	Access from	Link	Description	Fileformat	Comment
Global	**Latest three days** of Real-Time CAMS global **daily analysis and forecast** data	FTP	FTP dissemination service	??	??	
Global	CAMS Global **archived analysis and forecast** daily data	ECMWF	http://apps.ecmwf.int/datasets/data/cams-nrealtime/levtype=sfc/	??	GRIB/ NetCDF	
Global	CAMS global delayed-mode analyses and real-time forecasts of CO2 and CH4	Request only	copernicus-support@ecmwf.int	??	??	
Global	CAMS Global **archived reanalysis data**	ADS	CAMS global reanalysis (EAC4) 3-hourly data and CAMS global reanalysis (EAC4) monthly averaged fields	??	GRIB/ NetCDF	
Global	CAMS Global **archived** MACC Reanalysis	ECMWF	http://apps.ecmwf.int/datasets/data/macc-reanalysis/levtype=sfc/	??	GRIB/ NetCDF	
European	All CAMs regional **analysis and forecast** daily data	ADS	Atmosphere Data store (ADS) web interface	??	GRIB/ NetCDF	
European	CAMS regional reanalysis data	online	https://www.regional.atmosphere.copernicus.eu/index.php?category=data_access&subensemble=reanalysis_pro	??	NetCDF	

Eksempel: Udsnit af kategorisering af et dataset i Atmosphere kategorien

CAMS European Air quality	Variable	Model	Level	Start date(year-month-day)	End date(year-month-day)	Type	Time
https://ads.atmosphere.copernicus.eu/cdsap	Alder pollen	Ensamble median	0	earliest 2018-01-16	Now	Analysis	00:00
	Ammonia	CHIMERE	50			Forecast	01:00
	Birch pollen	EMEP	250				02:00
	Carbon Monoxide	LOTOS-EUROS	500				03:00
	Dust	MATCH	1000				04:00
	Grass pollen	MOCAGE	2000				05:00
	Nitrogen dioxide	SILAM	2000				06:00
	Nitrogen monoxide	EURAD-IM	3000				07:00
	Non-methane VOCs	DEHM	5000				08:00
	Olive pollen	GEM-AQ					09:00
	Ozone						10:00
	Particulate matter <2.5µm (PM2.5)						11:00
	Particulate matter <10µm (PM10)						12:00
	PM2.5, anthropognic fossil fuel carbon only						13:00
	PM2.5, anthropognic wood burning carbon only						14:00
	PM10, wildfires only						15:00
	Peroxyaxyl nitrates						16:00
	Ragwood pollen						17:00
	Secondary inorfanic aerosol						18:00
	Sulphur dioxide						19:00
							20:00
							21:00
							22:00
							23:00

Søgefunktion med Elasticsearch

Adgang til tekstdata

- Vi vil gerne automatisk kunne scrape data store for tekstdata som beskriver datasættene.
- Nogle hjemmesider viser indhold dynamisk ved hjælp af et javascript. Blandt andet atmosphere data storen.

https://ads.atmosphere.copernicus.eu/cdsapp#!/search?type=dataset

https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-global-reanalysis-eac4?tab=overv

Figure: HTML delen af linket ændres ikke - kun javascript delen ændres.

Metoder anvendt

- Typiske webcrawlers kan ikke scrape data fra hjemmesider baseret på javascripts.
- Webspider og scraper implementeret i Python.
- Pakker:
  - Scrapy[1]
  - Selenium[2] - langsommere, men virker - hurtig og effektiv, men virker ikke på javascriptspå javascripts
- Elasticsearch[3] - Open source enterprise search engine

[1] https://scrapy.org/
[2] https://www.selenium.dev/
[3] https://www.elastic.co/elastic-stack

Scrapy og Selenium

Vi har arbejdet med atmosphere data store og climate data store.

- https://ads.atmosphere.copernicus.eu/cdsapp#!/search?type=dataset
- https://cds.climate.copernicus.eu/cdsapp#!/search?type=dataset

For hvert dataset på disse hjemmesider bruger vi Scrapy og Selenium til at gemme

- url,
- titel,
- datarelateret tekst,
- parametre.

Dette bliver gemt i json format, som vi kan bruge i Elasticsearch.

Scrapy and Selenium

Eksempel på json format:

```
[
  {"Webpage": "https://ads.cop...",
   "Title": "CAMS global ...",
   "Description": ["EAC4 (ECMWF Atmospheric ... "],
   "Parameters":["2m dewpoint temperature", "2m temperature", ... ]
  }
  {"Webpage": "https://ads.cop...",
   "Title": "CAMS global ...",
   "Description": ["This data set contains ..."],
   "Parameters":["Carbon dioxide dry mole fraction", ... ]
  }
  ...
  ]
```

Elasticsearch

- Oprette en database i Elasticsearch
- Søgning

Elasticsearch

Oprette database

- Vi har oprettet en database i Elasticsearch.
- Hvert datasæt lader vi være et dokument i Elasticsearch.
- Dataet tilføjes til Elasticsearch med _bulk API:
- Dataet skal have et bestemt format for at vi kan bruge bulk:
  - Tilføjer et id og sætter de fire kategorier i samme tuborgklamme.

```
1 {"index": {"_id": "0"}}
2 { "Webpage": "https://ads.atmosphere.c...", "Title": "CAMS global reana...", "Description": [
    "EAC4 (ECMWF Atmospheric ..."], "Parameters": [ "2m dewpoint temperature", ...] }
3 {"index": {"_id": "1"}}
4 { "Webpage": "https://ads.atmosphere.co...", "Title": "CAMS global rean...", "Description": [
    "This data set contains ..." ], "Parameters": [ "Carbon dioxide dry mole fraction", "... ] }
```

Elasticsearch resultater

- Man kan søge på Elasticsearch eller direkte fra kommandolinjen ved hjælp af curl.

```
1  GET /copernicus3/_search?q=temperature
```

```
~$ curl -XGET "http://localhost:9200/copernicus3/_search?q=temperature&pretty"
```

Elasticsearch resultater


```
{
  "took" : 3,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 62,
      "relation" : "eq"
    },
    "max_score" : 3.9180772,
    "hits" : [
      {
        "_index" : "copernicus3",
        "_type" : "_doc",
        "_id" : "70",
        "_score" : 3.9180772,
        "_source" : {
          "Webpage" : "https://cds.climate.copernicus.eu/cdsapp#!/dataset/sis-temperature-statistics",
          "Title" : "Temperature statistics for Europe derived from climate projections",
          "Description" : [
            """This dataset contains temperature exposure statistics for Europe (e.g. percentiles) derived from the
               daily 2 metre mean, minimum and maximum air temperature for the entire year, winter (DJF: December
               -January-February) and summer (JJA: June-July-August). These statistics were derived within the C3S
               European Health service and are available for different future time periods and using different
               climate change scenarios.

Temperature percentiles are typically used in epidemiology and public health when defining health risk estimates
  and when looking at current and future health impacts, and they allow to identify a common threshold and
```

Elasticsearch resultater


```
{
  "took" : 3,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 62,
      "relation" : "eq"
    },
    "max_score" : 3.9180772,
    "hits" : [
      {
        "_index" : "copernicus3",
        "_type" : "_doc",
        "_id" : "70",
        "_score" : 3.9180772,
        "_source" : {
          "Webpage" : "https://cds.climate.copernicus.eu/cdsapp#!/dataset/sis-temperature-statistics",
          "Title" : "Temperature statistics for Europe derived from climate projections",
          "Description" : [
            "This dataset contains temperature exposure statistics for Europe (e.g. percentiles) derived from the da
ily 2 metre mean, minimum and maximum air temperature for the entire year, winter (DJF: December-January-February) a
```

Elasticsearch resultater

- Use case:
  - Find temperaturen i Tyskland

```
GET /copernicus3/_search?q=germany
```

```
1  GET /copernicus3/_search?q=(temperature OR germany)
```

```
1  GET /copernicus3/_search?q=(temperature AND germany)
```

Elasticsearch resultater (query = Germany)


```
{
  "took" : 3,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 1,
      "relation" : "eq"
    },
    "max_score" : 5.6566935,
    "hits" : [
      {
        "_index" : "copernicus3",
        "_type" : "_doc",
        "_id" : "30",
        "_score" : 5.6566935,
        "_source" : {
          "Webpage" : "https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-fire-burned-area",
          "Title" : "Fire burned area from 2001 to present derived from satellite observations",
          "Description" : [
            """The Burned Area products provide global information of total burned area (BA) at pixel and grid
               scale. The BA is identified with the date of first detection of the burned signal in the case of the
               pixel product, and with the total BA per grid cell in the case of the grid product. The products were
               obtained through the analysis of reflectance changes from medium resolution sensors (Terra MODIS,
               Sentinel-3 OLCI), supported by the use of MODIS thermal information. The burned area products also
               include information related to the land cover that has been burned, which has been extracted from the
               Copernicus Climate Change Service (C3S) land cover dataset, thus assuring consistency between the
               datasets.

The algorithms for BA retrieval were developed by the University of Alcala (Spain), and processed by Brockmann
  Consult GmbH (Germany). Different product versions are available. FireCCI v5.0cds and FireCCI v5.1cds were
  developed as part of the Fire ECV Climate Change Initiative Project (Fire CCI) and brokered to C3S, offering the
```

Elasticsearch resultater (query = Germany or temperature)


```
{
  "took" : 4,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 63,
      "relation" : "eq"
    },
    "max_score" : 5.6566935,
    "hits" : [
      {
        "_index" : "copernicus3",
        "_type" : "_doc",
        "_id" : "30",
        "_score" : 5.6566935,
        "_source" : {
          "Webpage" : "https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-fire-burned-area",
          "Title" : "Fire burned area from 2001 to present derived from satellite observations",
          "Description" : [
            """The Burned Area products provide global information of total burned area (BA) at pixel and grid scale. The BA is identified with the date of first detection of the burned signal in the case of the pixel product, and with the total BA per grid cell in the case of the grid product. The products were obtained through the analysis of reflectance changes from medium resolution sensors (Terra MODIS, Sentinel-3 OLCI), supported by the use of MODIS thermal information. The burned area products also include information related to the land cover that has been burned, which has been extracted from the Copernicus Climate Change Service (C3S) land cover dataset, thus assuring consistency between the datasets.

The algorithms for BA retrieval were developed by the University of Alcala (Spain), and processed by Brockmann Consult GmbH (Germany). Different product versions are available. FireCCI v5.0cds
```

Elasticsearch resultater (query = Germany and temperature)


```
{
  "took" : 1,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 0,
      "relation" : "eq"
    },
    "max_score" : null,
    "hits" : [ ]
  }
}
```

Elasticsearch resultater

- Grunden til at vi får respons ved at søge efter "Germany" er at ordet "Germany" indgår i teksten i forbindelse med en beskrivelse af hvor dataet er behandlet.
- Det har altså intet at gøre med målinger i Tyskland.
- Ved at søge "Germany or temperature" fås det førnævnte datasæt hvor Germany indgår i beskrivelsen og alle datasæt hvor temperature indgår.
- Det ses også, da vi får 0 resultater ved at søge på "Germany and temperature".

Hvad er der af begrænsninger i det hentede data?

- Use case:
  - Find temperaturen i Tyskland
    - Geografiske lokationer kan kun tilgås med logitude og latitude eller nord, syd, øst og vest indskrænkninger. Dermed er det ikke nemt at finde data for specifikke lande ved hjælp af søgning.
    - Parameteren temperatur indgår i 62 databaser under atmosphere og climate. Hvad er forskellen på disse? Hvilken en skal man bruge?
  - Det er lige nu også et problem at søge efter tidspunkter.
    - Det kan muligvis løses ved at scrape "Download Data", da der her er mulighed for at vælge tidspunkter.

Webpage Clustering

Webpage Clustering

- Indeling af data i clusters.
- Preprocessing
  - Tokenization (gør hvert ord i en sætning til et datapunkt).
  - Stemming og lemmasation (fjerner "ing", "ly", "s" og lignende så fx. predict og predicting bliver den samme værdi.)
  - Fjern stop words (fjerner ord som fx. the, and, is osv. og fjerener tegn som punktum, komma osv.).
- Generer features
- Afstandsmål baseret på features fx.
  - Euclidian
  - Jaccard
  - Cosine
- Implementation af en clustering algoritme fx. K-means.

Find Jaccard afstand mellem datasæt på CDS og ADS

- Vi vil undersøge om en lav Jaccard afstand mellem beskrivelserne af datasættene på ADS og CDS stemmer overens med at datasættene ligner hinanden.
- Vi bruger det data vi scrapet med Scrapy og Selenium. Dette er gemt i json format.
- Udsnit af data fra et website:


```
{"Webpage": "https://cds.climate.copernicus.eu/cdsapp#!/dataset/cems-glofas-seasonal-reforecast",
"Title": "Seasonal reforecasts of river discharge and related data from the Global Flood Awareness System",
"Description": ["This dataset provides a gridded modelled time series of river discharge forced with seasonal range meteorological reforecasts. The data is ..."],
"Parameters": ["River discharge in the last 24 hours", "Upstream area"]}
```

Preprocessing

- Vi konverterer tekst dataet for hvert datasæt til en streng. Vi har ekskluderet `"Webpage"` herfra.
- Eksempel på tekst data fra et datasæt:

```
"Seasonal reforecasts of river discharge and related data from the
Global Flood Awareness System ['This dataset provides a gridded
modelled time series of river discharge forced with seasonal range
meteorological reforecasts.  The data is ...']  ['River discharge in
the last 24 hours', 'Upstream area']"
```

Preprocessing
Tokenization

- Med tokenisation konverterer vi tekststrengen til en liste af ordene i strengen.
- Samtidig ændres alle bogstaver til små bogstaver.

```
['seasonal', 'reforecasts', 'of', 'river', 'discharge', 'and',
'related', 'data', 'from', 'the', 'global', 'flood', 'awareness',
'system', "'this", 'dataset', 'provides', 'a', 'gridded', 'modelled',
'time', 'series', 'of', 'river', 'discharge', 'forced', 'with',
'seasonal', 'range', 'meteorological', 'reforecasts', 'the', 'data',
'is', ..., "'river", 'discharge', 'in', 'the', 'last', 'hours',
"'upstream", 'area']
```

Preprocessing
Stemming

- Med stemming reducerer vi ord til deres stamme, således at ord med samme betydning i forskellige former bliver repræsenteret på samme måde i datasættet.
- Vi bruger Snowball stemming[4] gennem modulet nltk.[5]

```
['season', 'reforecast', 'of', 'river', 'discharg', 'and', 'relat',
'data', 'from', 'the', 'global', 'flood', 'awar', 'system', 'this',
'dataset', 'provid', 'a', 'grid', 'model', 'time', 'seri', 'of',
'river', 'discharg', 'forc', 'with', 'season', 'rang', 'meteorolog',
'reforecast', 'the', 'data', 'is', ..., 'river', 'discharg', 'in',
'the', 'last', 'hour', 'upstream', 'area']
```

[4] https://snowballstem.org/
[5] https://www.nltk.org/api/nltk.stem.html

Preprocessing
Removing stopwords

- Det sidste preprocessing step er at fjerne stopwords. Disse er de mest almindelige ord i et sprog.
- Der er ikke en universel liste af stopwords, og i denne implementering har vi anvendt listen af engelske stopwords i nltk modulet.

```
['season', 'reforecast', 'river', 'discharg', 'relat', 'data',
'global', 'flood', 'awar', 'system', 'dataset', 'provid', 'grid',
'model', 'time', 'seri', 'river', 'discharg', 'forc', 'season', 'rang',
'meteorolog', 'reforecast', 'data', ..., 'river', 'discharg', 'last',
'hour', 'upstream', 'area']
```

Jaccard afstand

- Vi konverterer de preprocessed lister til python sets.
- Dette reducerer mængden af data, da hvis et ord optræder flere gange i listen, vil det kun optræde en gang i det resulterende sæt.
- Vi har udregnet Jaccard afstanden mellem alle hjemmesiderne.

$$J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}, \tag{1}$$

hvor A og B er sæt for to hjemmesider og $J(A, B) \in [0, 1]$.

Udsnit af Jaccard matrix

	86	87	88	89	90	91	92	93	94
83	0.8675	0.912172	0.912219	0.911579	0.851658	0.852941	0.851138	0.876259	0.867424
84	0.864486	0.923592	0.923077	0.920981	0.874593	0.875606	0.868637	0.9	0.869176
85	0.868966	0.916667	0.916041	0.913934	0.859247	0.86039	0.853135	0.88587	0.866548
86	0	0.911765	0.915483	0.914865	0.868254	0.867508	0.864217	0.882353	0.858392
87	0.911765	0	0.0487106	0.0762174	0.731488	0.727147	0.750171	0.718218	0.859807
88	0.915483	0.0487106	0	0.0306122	0.726953	0.726124	0.74601	0.717131	0.858355
89	0.914865	0.0762174	0.0306122	0	0.742361	0.740638	0.730526	0.715803	0.853307
90	0.868254	0.731488	0.726953	0.742361	0	0.0403397	0.115226	0.769504	0.827869
91	0.867508	0.727147	0.726124	0.740638	0.0403397	0	0.139394	0.766509	0.827446
92	0.864217	0.750171	0.74601	0.730526	0.115226	0.139394	0	0.774764	0.816044
93	0.882353	0.718218	0.717131	0.715803	0.769504	0.766509	0.774764	0	0.815217
94	0.858392	0.859807	0.858355	0.853307	0.827869	0.827446	0.816044	0.815217	0
95	0.885417	0.942623	0.942401	0.940278	0.89913	0.9	0.891037	0.888235	0.863095

Figure: Et udsnit af Jarccard Matricen

Resultater

- Lav afstand svarer til overlap af hjemmesiderne.
- Eksempel:
  - Original hjemmeside: Link
  - Tætteste hjemmeside: Link, Score: 0.03663793103448276
  - Femte bedste: Link, Score: 0.7316047652417659
- Der ses en klar relation mellem beskrivelsen i alle tre datasæt, men det femte bedste datasæt har nogle helt andre parametre, hvilket giver udslaget i score.

Kode og data

- Scripts anvendt i forbindelse med arbejdet findes i vores Github repository.
- Her findes også en readme angående bl.a. installation af pakker.